Anomaly Detection by Robust Statistics

Peter J. Rousseeuw and Mia Hubert

October 14, 2017

Abstract

Real data often contain anomalous cases, also known as outliers. These may spoil the resulting analysis but they may also contain valuable information. In either case, the ability to detect such anomalies is essential. A useful tool for this purpose is robust statistics, which aims to detect the outliers by first fitting the majority of the data and then flagging data points that deviate from it. We present an overview of several robust methods and the resulting graphical outlier detection tools. We discuss robust procedures for univariate, low-dimensional, and high-dimensional data, such as estimating location and scatter, linear regression, principal component analysis, classification, clustering, and functional data analysis. Also the challenging new topic of cellwise outliers is introduced.

Department of Mathematics, KU Leuven, Celestijnenlaan 200B, BE-3001 Leuven, Belgium.

INTRODUCTION

In real data sets it often happens that some cases behave differently from the majority. Such data points are called *anomalies* in machine learning, and *outliers* in statistics. Outliers may be caused by errors, but they could also have been recorded under exceptional circumstances, or belong to another population. It is very important to be able to detect anomalous cases, which may (a) have a harmful effect on the conclusions drawn from the data, or (b) contain valuable nuggets of information.

In practice one often tries to detect outliers using diagnostics starting from a classical fitting method. However, classical methods can be affected by outliers so strongly that the resulting fitted model may not allow to detect the deviating observations. This is called the *masking* effect. Additionally, some good data points might even appear to be outliers, which is known as *swamping*. To avoid these effects, the goal of robust statistics is to find a fit which is close to the fit we would have found without the outliers. We can then identify the outliers by their large 'deviation' (e.g. its distance or residual) from that robust fit.

First we describe some robust procedures for detecting anomalies in univariate location and scale, as well as in multivariate data and in the linear regression setting. For more details on this part see[1–3]. Next we discuss principal component analysis (PCA) and some available robust methods for classification, clustering, and functional data analysis. Finally we introduce the emerging research topic of detecting cellwise anomalies.

ESTIMATING UNIVARIATE LOCATION AND SCALE

As an example of univariate data, suppose we have five measurements of a length:

$$6.27, \quad 6.34, \quad 6.25, \quad 6.31, \quad 6.28 \tag{1}$$

and we want to estimate its true value. For this, one usually computes the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ which in this case equals $\bar{x} = (6.27 + 6.34 + 6.25 + 6.31 + 6.28)/5 = 6.29$. Let us now suppose that the fourth measurement has been recorded wrongly and the data become

$$6.27, \quad 6.34, \quad 6.25, \quad 63.1, \quad 6.28\,. \tag{2}$$

In this case we obtain $\bar{x} = 17.65$, which is far off. Alternatively, we could also compute the *median* of these data. For this we sort the observations in (2) from smallest to largest:

$$6.25 \leqslant 6.27 \leqslant 6.28 \leqslant 6.34 \leqslant 63.10 \ .$$

The median is the middle value, here yielding 6.28, which is still reasonable. We say that the median is more robust against an outlier.

More generally, the location-scale model states that the n univariate observations x_i are independent and identically distributed (i.i.d.) with distribution function $F((x-\mu)/\sigma)$ where F is known. Typically F is the standard gaussian distribution function Φ. We then want to find estimates for the unknown center μ and the unknown scale parameter σ.

The classical estimate of location is the mean. As we saw above, the mean is very sensitive to aberrant values among the n observations. In general, replacing even a single observation by a very large value can change the mean completely. We say that the *breakdown value*[4,5] of the sample mean is $1/n$, so it becomes 0% for large n. In general, the breakdown value is the smallest proportion of observations in the data set that need to be replaced to carry the estimate arbitrarily far away. A breakdown value of 0% is thus the worst possible. See[6] for precise definitions and extensions. The robustness of an estimator is also measured by its *influence function*[7] which measures the effect of a single outlier. The influence function of the mean is unbounded, which again illustrates that the mean is not robust.

For the general definition of the median, we denote the ith ordered observation as $x_{(i)}$. The median is defined as $x_{((n+1)/2)}$ if n is odd and $(x_{(n/2)} + x_{(n/2+1)})/2$ if n is even. Its breakdown value is about 50%, meaning that the median can resist almost 50% of outliers. This is the best possible breakdown value since the clean data need to be in the majority.

The situation for the scale parameter σ is similar. The classical estimator is the *standard deviation* $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$. Since a single outlier can already make s arbitrarily large, its breakdown value is 0%. For instance, for the clean data (1) above we have $s = 0.035$, whereas for the data (2) with the outlier we obtain $s = 25.41$!

A robust measure of scale is the **m**edian of all **a**bsolute **d**eviations from the median (MAD), given by

$$\text{MAD} = 1.4826 \operatorname*{median}_{i=1,\ldots,n} |x_i - \operatorname*{median}_{j=1,\ldots,n}(x_j)| \ . \tag{3}$$

The constant 1.4826 is a correction factor which makes the MAD consistent at gaussian distributions. The MAD of (2) is the same as that of (1), namely 0.044. We can also use the Q_n estimator[8], defined as

$$Q_n = 2.2219\,\{|x_i - x_j|; i < j\}_{(k)}$$

with $k = \binom{h}{2} \approx \binom{n}{2}/4$ and $h = \lfloor \frac{n}{2} \rfloor + 1$. Here $\lfloor \ldots \rfloor$ rounds down to the nearest integer. This scale estimator is thus the first quartile of all pairwise distances between two data points. The breakdown value of both the MAD and the Q_n estimator is 50%.

Also the (normalized) interquartile range (IQR) can be used, given by IQR $= 0.7413(Q_3 - Q_1)$ where $Q_1 = x_{\lfloor n/4 \rfloor}$ is the first quartile of the data and $Q_3 = x_{\lceil 3n/4 \rceil}$ is the third quartile. The IQR has a simple expression but its breakdown value is only 25%, so it is less robust than the MAD and Q_n.

The robustness of the median (and the MAD) comes at a price: at the gaussian model it is less efficient than the mean. Many robust procedures have been proposed that strike a balance between robustness and efficiency, such as location M-estimators[9]. They are defined implicitly as the solution of the equation

$$\sum_{i=1}^{n} \psi\Big(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\Big) = 0 \tag{4}$$

for a given real function ψ. The denominator $\hat{\sigma}$ is an initial robust scale estimate such as Q_n. A solution $\hat{\mu}$ to (4) can be found by an iterative algorithm, starting from the initial location estimate $\hat{\mu}^{(0)} = \text{median}_i(x_i)$. Popular choices for ψ are the Huber function $\psi(x) = x \min(1, c/|x|)$ and Tukey's bisquare function $\psi(x) = x(1-(x/c)^2)^2 I(|x| \leqslant c)$. These M-estimators contain a tuning parameter c which needs to be chosen in advance. Also M-estimators for the scale parameter σ exist.

People often use rules to detect outliers. The classical rule is based on the z-scores of the observations, given by

$$z_i = \frac{x_i - \bar{x}}{s} \tag{5}$$

where s is the standard deviation of the data. More precisely, the rule flags x_i as outlying if $|z_i|$ exceeds 2.5, say. But in the above example (2) with the outlier, the z-scores are

$$-0.45, \quad -0.45, \quad -0.45, \quad 1.79, \quad -0.45$$

so none of them attains 2.5. The largest value is only 1.79, which is quite similar to the largest z-score for the clean data (1), which equals 1.41. The z-score of the outlier is small because it subtracts the nonrobust mean (which was drawn toward the outlier) and because it divides by the nonrobust standard deviation (which the outlier has made much larger than in the clean data). Plugging in robust estimators of location and scale such as the median and the MAD yields the robust scores

$$\frac{x_i - \text{median}_j(x_j)}{\text{MAD}_j(x_j)} \tag{6}$$

which yield a much more reliable outlier detection tool. Indeed, in the contaminated example (2) the robust scores are

$$-0.22, \quad 1.35, \quad -0.67, \quad 1277.5, \quad 0.0$$

where that of the outlier greatly exceeds the 2.5 cutoff.

Also Tukey's boxplot is often used to pinpoint possible outliers. In this plot a box is drawn from the first quartile Q_1 of the data to the third quartile Q_3 . Points outside the interval $[Q_1 - 1.5\ \text{IQR}, Q_3 + 1.5\ \text{IQR}]$, called the *fence*, are traditionally marked as outliers. Note that the boxplot assumes symmetry, since we add the same amount to Q_3 as what we subtract from Q_1. At asymmetric distributions the usual boxplot typically flags many regular data points as outliers. The skewness-adjusted boxplot[10] corrects for this by using a robust measure of skewness[11] in determining the fence.

MULTIVARIATE LOCATION AND COVARIANCE ESTIMATION

From now on we assume that the data are d-dimensional and are stored in an $n \times d$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^T$ the ith data point. Classical measures of location and scatter are given by the empirical mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ and the empirical covariance matrix $\boldsymbol{S_X} = \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T/(n-1)$. As in the univariate case, both classical estimators have a breakdown value of 0%, that is, a small fraction of outliers can completely ruin them.

As an illustration we consider a bivariate dataset (from page 59 in[2]) containing the logarithms of body weight and brain weight of 28 animal species, with scatterplot in Figure 1. Any point $\boldsymbol{x}$ has a so-called Mahalanobis distance (or 'generalized distance')

$$\mathrm{MD}(\boldsymbol{x}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sqrt{(\boldsymbol{x} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}})} \tag{7}$$

to the mean $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$, taking the covariance matrix $\hat{\boldsymbol{\Sigma}} = \boldsymbol{S_X}$ into account. The MD is constant on ellipsoids. The so-called 97.5% tolerance ellipsoid is given by $MD(\boldsymbol{x}) \leqslant \sqrt{\chi^2_{d,0.975}}$ where $\chi^2_{d,0.975}$ is the 0.975 quantile of the chi-squared distribution with d degrees of freedom. In this bivariate example $d = 2$, and the resulting ellipse is drawn in red. We see that it is inflated in the direction of the three outliers 6, 16, and 26 which are dinosaurs having low brain weight and high body weight. As a result these data points fall near the boundary of the tolerance ellipse, i.e. their $MD(\boldsymbol{x}_i)$ are not very high.

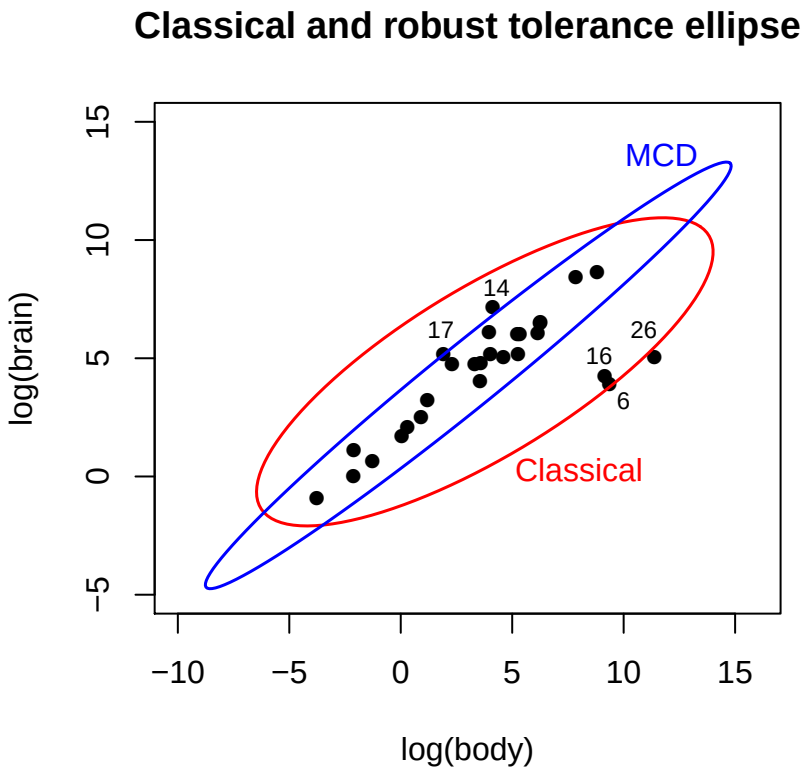

Figure 1: Animal data: tolerance ellipse of the classical mean and covariance matrix (red), and that of the robust location and scatter matrix (blue).

Alternatively we can compute robust estimates of location and scatter (covariance), for instance by the *Minimum Covariance Determinant* (MCD) method[12,13]. The MCD looks for those h observations in the data set (where the number h is given by the user) whose classical covariance matrix has the lowest possible determinant. The MCD estimate of location $\hat{\boldsymbol{\mu}}$ is then the average of these h points, whereas the MCD estimate of scatter $\hat{\boldsymbol{\Sigma}}$ is their covariance matrix, multiplied by a consistency factor. (By default this is then followed by a reweighting step to improve efficiency at gaussian data.) Instead of Mahalanobis distances we can then compute robust distances, again given by (7) but now with the robust estimates $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$. This yields the robust tolerance ellipse shown in blue in Figure 1. This ellipse exposes the three dinosaurs, and we see two species near the upper boundary, 17 (rhesus monkey) and 14 (human).

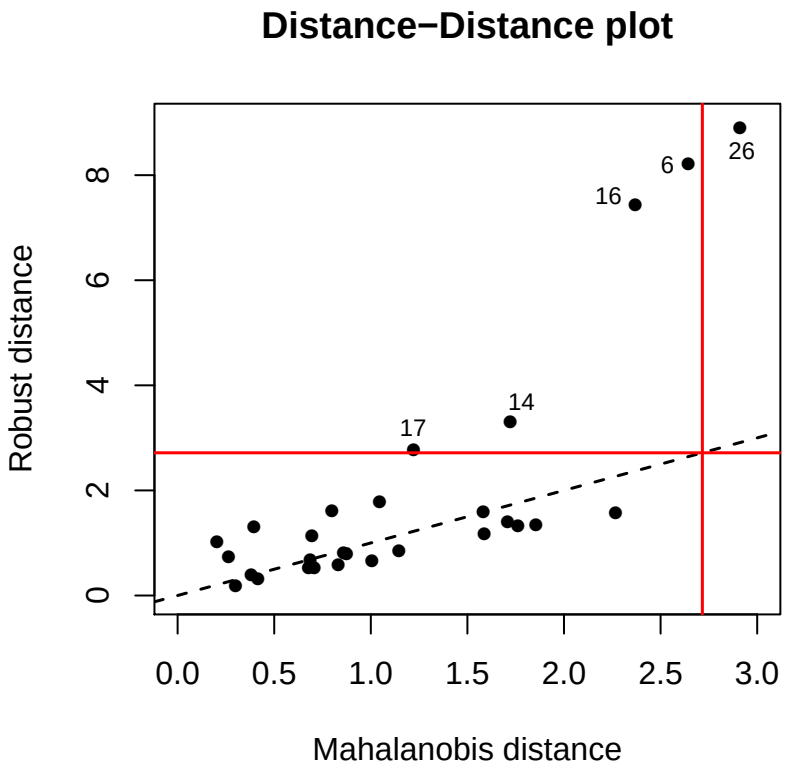

Figure 2: Animal data: Robust distance versus classical Mahalanobis distance.

In dimension $d = 4$ or higher it becomes infeasible to visualize the tolerance ellipsoid, but we still have the distances. The *distance-distance plot* (DD-plot) in Figure 2 shows the

robust distance RD($\boldsymbol{x}_i$) of each data point versus its classical Mahalanobis distance MD($\boldsymbol{x}_i$) . The horizontal and vertical cutoff lines are at $\sqrt{\chi^2_{d,0.975}}$ and the dashed line is where classical and robust distances coincide. We see that the $RD(\boldsymbol{x}_i)$ flag all the outliers in this dataset, while the $MD(\boldsymbol{x}_i)$ don't. For a dataset in which they are very similar we can trust classical statistical methods, but when they differ much (like here) the DD-plot detects the outlying data points. This does not imply we should somehow delete them, but rather that they should be investigated and understood. Outliers are not necessarily 'errors': they can also correspond to unusual circumstances or be members of a different population.

The MCD estimator, as well as its weighted version, has a bounded influence function and breakdown value $(n-h+1)/n$, hence the number h determines the robustness of the estimator. The MCD has its highest possible breakdown value when $h = \lfloor (n+p+1)/2 \rfloor$. When a large proportion of contamination is expected, h should thus be chosen close to $0.5n$. Otherwise an intermediate value for h, such as $0.75n$, is recommended to obtain a higher finite-sample efficiency. Reference[14] gives a more detailed overview of the MCD estimator and its properties.

The computation of the MCD estimator is non-trivial and naively requires an exhaustive investigation of all h-subsets out of n. Fortunately a much faster algorithm was constructed, called FastMCD[15]. It starts by randomly drawing many $p+1$ observations from the data set. Based on these subsets, h-subsets are obtained by means of so-called C-steps (see[15] for details). More recently an even faster algorithm called DetMCD was devised[16] which carries out a deterministic computation instead of random sampling.

The MCD assumes that $n > d$, so there must be more data points than dimensions, and it works best when $n > 5d$. When there are more than, say, 20 dimensions and/or $d \geqslant n$ other methods are needed. One is to compute robust principal components as described in a section below. Another is to use the *minimum regularized covariance determinant* (MRCD) method[17]. This approach minimizes $\det\{\rho\boldsymbol{T}+(1-\rho)\boldsymbol{S}_H\}$ where $\boldsymbol{T}$ is a positive definite target matrix and $\boldsymbol{S}_H$ is the covariance matrix of a subset H with h data points. The combined matrix is always positive definite, whereas $\det\{\boldsymbol{S}_H\} = 0$ when $d \geqslant n$.

Many other robust estimators of location and scatter have been presented in the literature. The first such estimator was proposed by Stahel[18] and Donoho[19] (see also[21]). They defined

the so-called Stahel-Donoho outlyingness of a data point $\boldsymbol{x}_i$ as

$$\mathrm{outl}(\boldsymbol{x}_i) = \max_{\boldsymbol{u}} \frac{|\boldsymbol{x}_i^T\boldsymbol{u} - \mathrm{median}_{j=1,\ldots,n}(\boldsymbol{x}_j^T\boldsymbol{u})|}{\mathrm{MAD}_{j=1,\ldots,n}(\boldsymbol{x}_j^T\boldsymbol{u})} \tag{8}$$

where the maximum is over all directions (i.e., all d-dimensional unit length vectors $\boldsymbol{u}$), and $\boldsymbol{x}_j^T\boldsymbol{u}$ is the projection of $\boldsymbol{x}_j$ on the direction $\boldsymbol{u}$. In each direction this uses the robust z-scores (6). Recently a version of (8) suitable for skewed distributions was proposed[20].

Multivariate M-estimators[22] have a low breakdown value due to possible implosion of the estimated scatter matrix. More recent robust estimators of multivariate location and scatter with high breakdown value include S-estimators[2,23], MM-estimators[24], and the OGK estimator[25].

LINEAR REGRESSION

The multiple linear regression model assumes that there are d 'explanatory' x-variables as well as a response variable y which can be approximated by a linear combination of the x-variables. More precisely, the model says that for all data points $(\boldsymbol{x}_i, y_i)$ it holds that

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + \varepsilon_i \qquad i = 1, \ldots, n \tag{9}$$

where the errors ε_i are assumed to be independent and identically distributed with zero mean and constant variance σ^2. Applying a regression estimator to the data yields $d+1$ regression coefficients, combined as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_d)^T$. The residual r_i of case i is defined as the difference between the observed response y_i and its estimated value $\hat{y}_i$.

The classical least squares (LS) method to estimate $\boldsymbol{\beta}$ minimizes the sum of the squared residuals. It is popular because it allows to compute the regression estimates explicitly, and it is optimal if the errors have a gaussian distribution. Unfortunately LS is extremely sensitive to outliers, i.e. data points that do not obey the linear pattern formed by the majority of the data.

For instance, Figure 3 shows the Hertzsprung-Russell diagram of the star cluster CYG OB1, containing 47 stars. The x-coordinate of each star is the logarithm of its surface temperature, and the y-coordinate is the logarithm of its light intensity. Most of the stars belong

to the so-called main sequence, whereas 11, 20, 30, 34 are giant stars and 7 is intermediate. The least squares line is shown in red, and has a negative slope although the main sequence slopes upward. It has been pulled away by the leverage exerted by the four giant stars. As an unfortunate side effect, the giant stars do not have larger absolute residuals than some of the main sequence stars, so only looking at residuals would not allow to detect them.

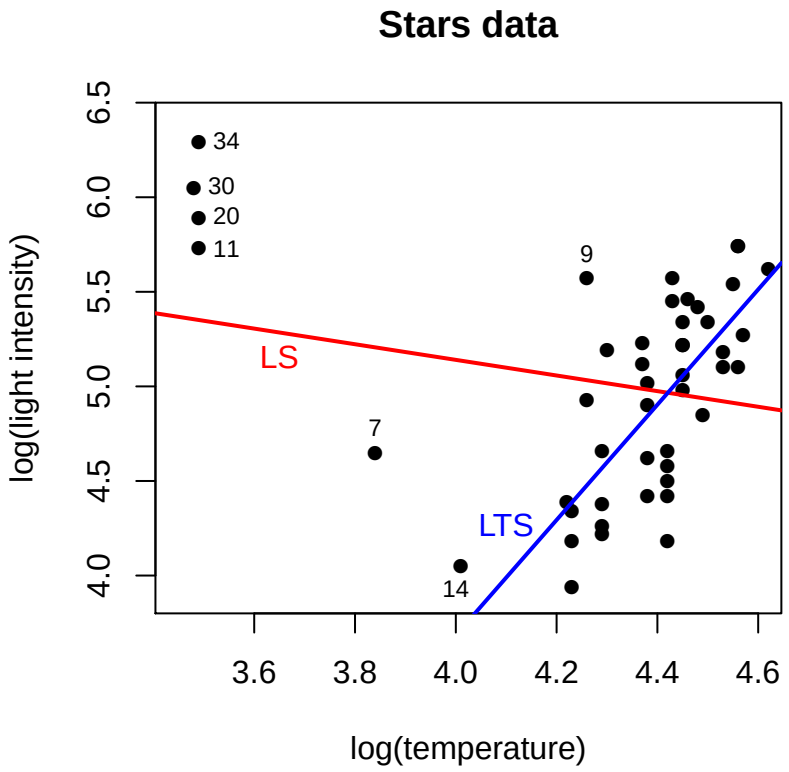

Figure 3: Stars data: Classical least squares line (red) and robust line (blue).

The blue line on the other hand is the result of a robust regression method, the *Least Trimmed Squares* (LTS) estimator proposed by Rousseeuw[12]. The LTS is given by

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{h} (r^2)_{(i)} \tag{10}$$

where $(r^2)_{(1)} \leqslant (r^2)_{(2)} \leqslant \cdots \leqslant (r^2)_{(n)}$ are the ordered squared residuals. (They are first squared, and then ordered.) By not adding *all* the squared residuals the LTS makes it possible to fit the majority of the data, whereas the outliers can have large residuals. In Figure 3 the blue line indeed fits the main sequence stars, and stays far from the four giant

stars so the latter will have large residuals from that line. (Note that the giant stars are not 'errors' but correct observations of members of a different population.)

The value h in (10) plays the same role as in the MCD estimator. For $h \approx n/2$ we find a breakdown value of 50%, whereas for larger h we obtain roughly $(n-h)/n$. A fast algorithm for the LTS estimator (FAST-LTS) has been developed[26]. The scale of the errors σ can be estimated by $\hat{\sigma}^2_{\mathrm{LTS}} = c^2_{h,n} \sum_{i=1}^{h} (r^2)_{(i)}/h$ where r_i are the residuals from the LTS fit, and $c_{h,n}$ is a constant that makes $\hat{\sigma}_{\mathrm{LTS}}$ consistent at gaussian error distributions, as described in[27]. We can then identify outliers by their large standardized LTS residuals $r_i/\hat{\sigma}_{\mathrm{LTS}}$. We can also use the standardized LTS residuals to assign a weight to every observation. The weighted LS estimator with these LTS weights inherits the nice robustness properties of LTS, but is more efficient and yields all the usual inferential output such as t-statistics, F-statistics, an R^2 statistic, and the corresponding p-values. Alternatively, inference for LTS can be based on the fast robust bootstrap proposed in[28,29].

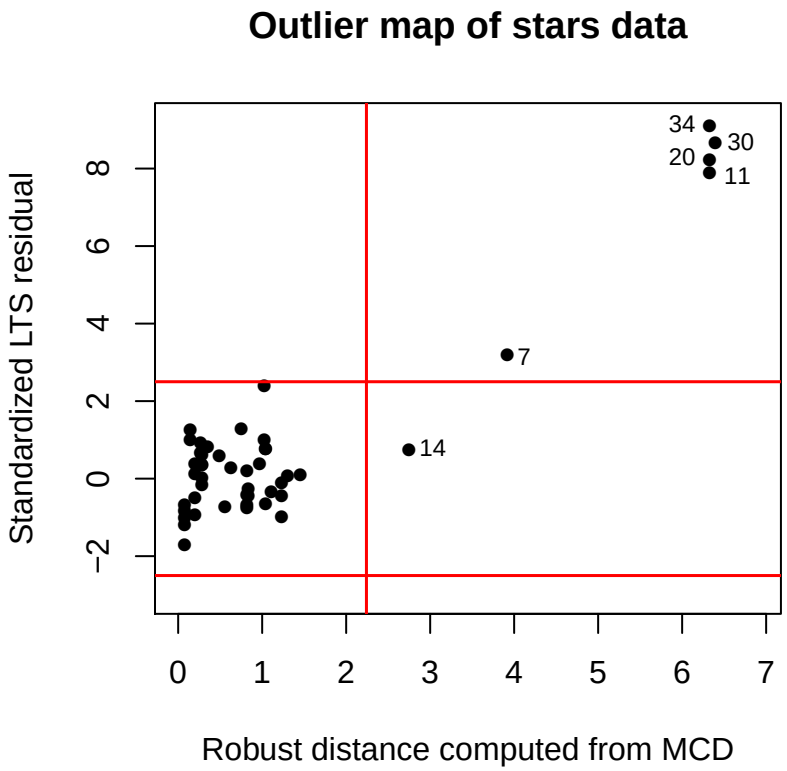

Figure 4: Stars data: Standardized robust residuals of y versus robust distances of x.

In most situations we have more than one explanatory variable, and for dimension $d = 3$ and higher it is no longer possible to perceive the linear patterns by eye. It is in those cases that robust regression becomes the most useful. To flag and interpret the outliers we can use the *outlier map* of[30] which plots the standardized LTS residuals versus robust distances (7) based on (for instance) the MCD estimator applied to the x-variables only. Figure 4 is the outlier map of the stars data. The tolerance band on the standardized LTS residuals is given by the horizontal lines at 2.5 and -2.5 , and the vertical line corresponds to the cutoff value $\sqrt{\chi^2_{d,0.975}}$ on the robust distances of the $\boldsymbol{x}_i$. Data points $(\boldsymbol{x}_i, y_i)$ whose residuals fall outside the horizontal tolerance band are called *regression outliers*. On the other hand, data points $(\boldsymbol{x}_i, y_i)$ whose robust distance $RD(\boldsymbol{x}_i)$ exceeds the cutoff are called *leverage points*, irrespective of their response y_i. So, the outlier map diagnoses 4 types of data points. Those with small $|r_i|$ and small $RD(\boldsymbol{x}_i)$ are considered *regular observations*, and most points in Figure 4 fall in that rectangle. Those with large residuals r_i (positive or negative) and small $RD(\boldsymbol{x}_i)$ are called *vertical outliers* (there are none in this figure). Those with small $|r_i|$ but large $RD(\boldsymbol{x}_i)$ (like point 14) are called *good leverage points* because they improve the accuracy of the fit. And finally, regression outliers that are also leverage points are called *bad leverage points*, like the 4 giant stars in this example. Note that the outlier map permits nuanced statements, for instance point 7 is a leverage point but only slightly bad.

The main benefit of the outlier map is when the data has more dimensions. For instance, the stackloss data[31] is a benchmark data set with 21 points with $d = 3$ explanatory variables, an intercept term and a response variable y_i. We cannot easily interpret such 4-dimensional data, but we can still look at the outlier map in the right panel of Figure 5. We see that 4 is a vertical outlier, 1, 3, and 21 are bad leverage points, and 2 is a good leverage point. Note that the left panel of Figure 5 does not flag any of these points because it uses the classical LS residuals and the classical distances $MD(\boldsymbol{x}_i)$, both of which tend to mask atypical points.

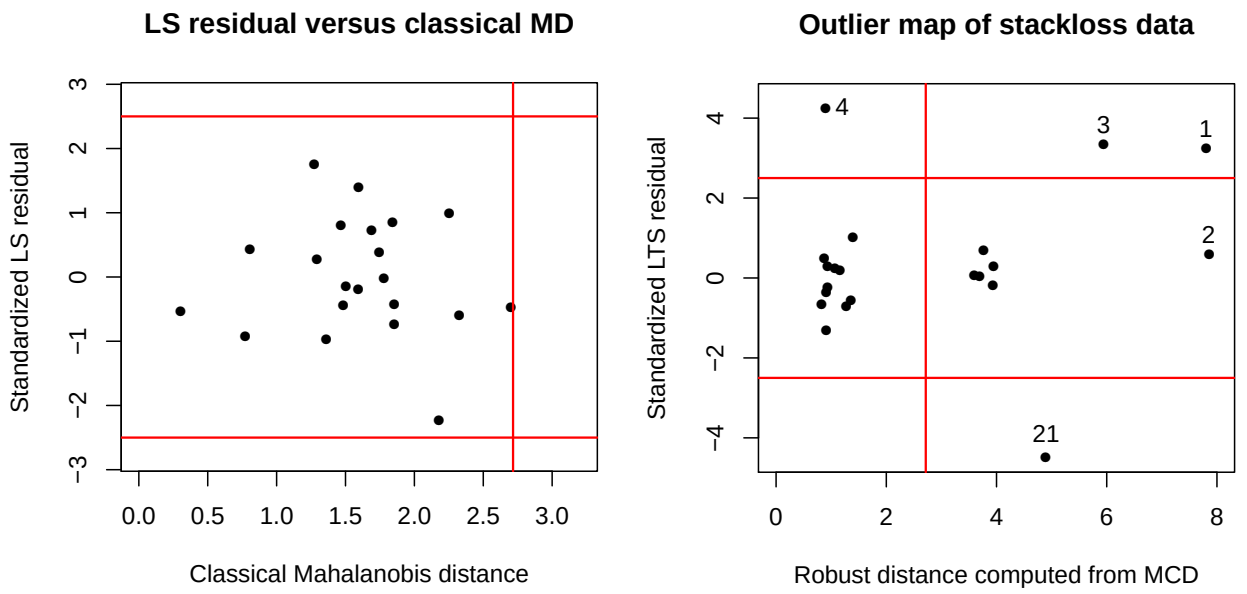

Figure 5: Stackloss data: (left) standardized nonrobust LS residuals of y versus nonrobust distances of x; (right) same with robust residuals and robust distances.

It should be stressed that flagging atypical points with the outlier map (as in the right panel of Figure 5) is not the end of the analysis, but rather a new start. The next step should be to try to figure out why these points are atypical and/or to improve the model by things like data transformation, model selection, higher order terms, etc. For variance selection in robust regression see[32]. When the dimension is very high one needs to resort to sparse methods, for instance by penalization. The first sparse methods for robust regression were developed in[33,34].

Historically, the earliest attempts at robust regression were least absolute deviations (LAD, also called L^1), M-estimators[35], R-estimators[36], and L-estimators[37]. The breakdown value of all these methods is 0% because of their vulnerability to bad leverage points. Generalized M-estimators (GM-estimators)[7] were the first to attain a positive breakdown value, which unfortunately still went down to zero for increasing p.

The low finite-sample efficiency of LTS can be improved by replacing its objective function by a more efficient scale estimator applied to the residuals r_i. This approach has led to the introduction of high-breakdown regression S-estimators[38] and MM-estimators[39].

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a popular dimension reduction method. It tries to explain the covariance structure of the data by means of a (hopefully small) number of components. These components are linear combinations of the original variables, and often allow for an interpretation and a better understanding of the different sources of variation. PCA is often the first step of the data analysis, followed by other multivariate techniques.

In the classical approach, the first principal component corresponds to the direction in which the projected data points have the largest variance. The second component is then taken orthogonal to the first and must again maximize the variance of the data points projected on it (subject to the orthogonality constraint). Continuing in this way produces all the principal components. It turns out that the classical principal components correspond to the eigenvectors of the empirical covariance matrix. Unfortunately, both the classical variance (which is being maximized) and the classical covariance matrix (which is being decomposed) are very sensitive to anomalous observations. Consequently, the first components from classical PCA are often attracted towards outlying points, and may not capture the variation of the regular observations.

A first group of robust PCA methods is obtained by replacing the classical covariance matrix by a robust covariance estimator, such as the weighted MCD estimator or MM-estimators[40,41]. Unfortunately the use of these covariance estimators is limited to small to moderate dimensions since they are not defined when $d \geqslant n$.

A second approach to robust PCA uses *Projection Pursuit* techniques. These methods maximize a robust measure of spread to obtain consecutive directions on which the data points are projected, see[42,43].

The ROBPCA[44] approach is a hybrid, which combines ideas of projection pursuit and robust covariance estimation. The projection pursuit part is used for the initial dimension reduction. Some ideas based on the MCD estimator are then applied to this lower-dimensional data space.

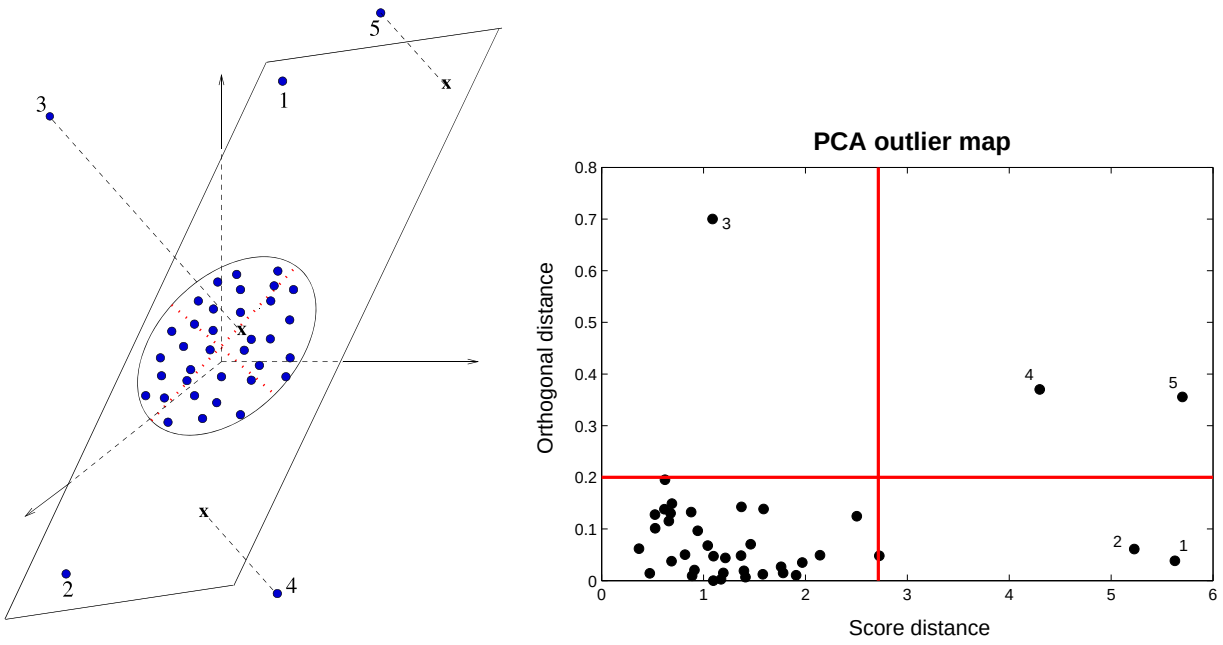

Figure 6: Illustration of PCA: (left) types of outliers; (right) outlier map: plot of orthogonal distances versus score distances.

In order to diagnose outliers we can draw an outlier map for PCA [44], similar to the outlier map for regression in the previous section. A stylized example of such a PCA outlier map is shown in the right panel of Figure 6, which corresponds to the three-dimensional data in the left panel which is fitted by two principal components. On the vertical axis of the PCA outlier map we find the *orthogonal distance* of each data point to the PCA subspace. This is just the Euclidean distance of the data point to its projection. The orthogonal distance is highest for the points 3, 4, and 5 in the example. On the horizontal axis we see the *score distance* of each data point, which is just the robust distance (7) of its projection relative to all the projected data points. The score distance is rather high for the points 1, 2, 4, and 5 in the figure.

By combining both distance measures the outlier map allows to distinguish between four types of data points. *Regular observations* have both a small orthogonal distance and a small score distance. Points with a high score distance but a small orthogonal distance, such as points 1 and 2 in Figure 6, are called *good leverage points* as they can improve the accuracy of the fitted PCA subspace. *Orthogonal outliers* have a large orthogonal distance but a small

score distance, like point 3. *Bad leverage points* have both a large orthogonal distance and a large score distance, like points 4 and 5. They lie far from the space spanned by the robust principal components, and after projection on that space they lie far from most of the other projected data. They are called 'bad' because they typically they have a large influence on classical PCA, as the eigenvectors will be tilted towards them.

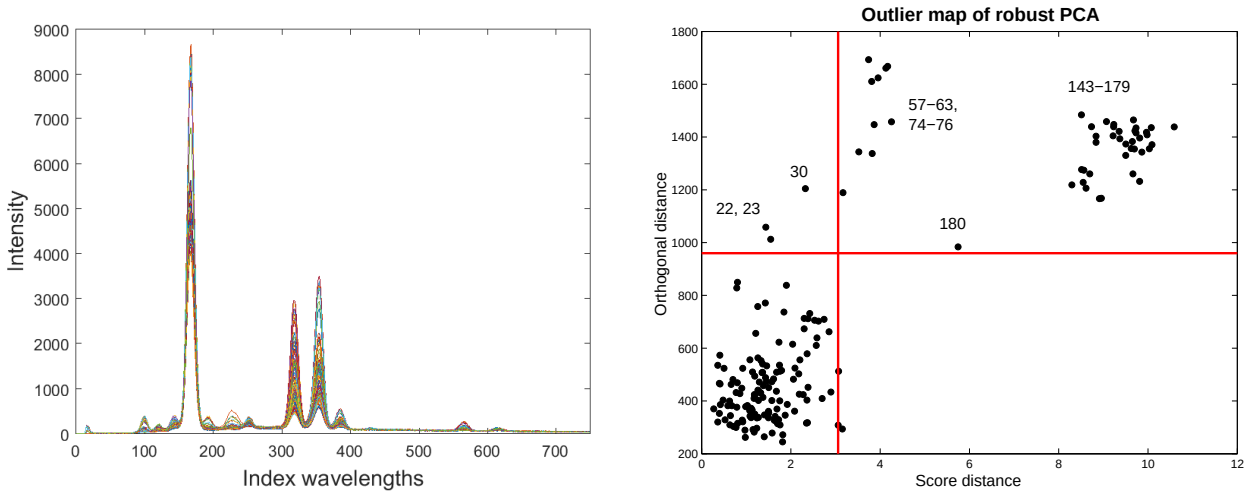

Figure 7: Glass data: (left) spectra; (right) outlier map.

As a real example we take the glass data[45] consisting of spectra of 180 archaeological glass vessels from the 16th–17th centuries. They have 750 wavelengths so $d > n$. The spectra are shown in Figure 7 with their outlier map based on ROBPCA, which clearly indicates a substantial number of bad leverage points and several orthogonal outliers. An analogous plot based on classical PCA (not shown) did not reveal the outliers, because they tilted the PCA subspace toward them. Also the plots of the first few principal components looked quite different.

Other proposals for robust PCA include spherical PCA[46] which first projects the data onto a sphere with a robust center, and then applies PCA to these projected data. To obtain sparse loadings, a robust sparse PCA method is proposed in[47]. When linear models are not appropriate, one may use support vector machines (SVM) which are powerful tools for handling nonlinear structures[48]. A kernelized version of ROBPCA (KROBPCA) is introduced in[49]. For a review of robust versions of principal component regression and partial least squares see[1].

OTHER MODELS

Classification

The goal of classification, also known as discriminant analysis or supervised learning, is to obtain rules that describe the separation between known groups G_j of d-dimensional data points, with an eye toward assigning new data points to one of the groups. We write p_j for the membership probability, i.e. the probability for any observation to come from G_j.

For low-dimensional data, a popular classification rule results from maximizing the Bayes posterior probability. At gaussian distributions this yields quadratic discriminant analysis (QDA), i.e. choosing the j for which $\boldsymbol{x}$ has the highest quadratic score $d_j^Q(\boldsymbol{x})$ given by

$$d_j^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) + \ln(p_j)\ . \tag{11}$$

When all the covariance matrices are assumed to be equal, these scores can be simplified to

$$d_j^L(\boldsymbol{x}) = \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln(p_j) \tag{12}$$

where $\boldsymbol{\Sigma}$ is the common covariance matrix, yielding linear discriminant analysis (LDA). Robust classification rules can be obtained by replacing the classical covariance matrices by robust alternatives such as the MCD estimator or S-estimators, as in [50–53].

When the data are high-dimensional, this approach cannot be applied because the robust covariance estimators are no longer computable. One approach is to first apply robust PCA to the entire data set. Alternatively, one can also apply a PCA method to each group separately. This is the idea behind the SIMCA (Soft Independent Modeling of Class Analogy) method [54]. A robustification of SIMCA is obtained by first applying robust PCA to each group, and then constructing a classification rule for new observations based on their orthogonal distance to each subspace and their score distance within each subspace [55].

An SVM classifier with an unbounded kernel, e.g. a linear kernel, is not robust and suffers the same problems as traditional linear classifiers. But when a bounded kernel is used, the resulting non-linear SVM classification handles outliers quite well [48]. As an alternative, one can apply KROBPCA combined with LDA to the scores [49].

Clustering

Cluster analysis (also known as unsupervised learning) is an important methodology when handling large data sets. It searches for homogeneous groups in the data, which afterwards may be analyzed separately. Partitioning (non-hierarchical) clustering methods search for the best clustering in k groups.

For spherical clusters, the most popular method is k-means which minimizes the sum of the squared Euclidean distances of the observations to the mean of their group[56]. This method is not robust as it uses averages. To overcome this problem, one of the first robust proposals was the Partitioning Around Medoids method[57]. It searches for k observations (called medoids) such that the sum of the unsquared distances of the observations to the medoid of their group is minimized. The CLARA algorithm[57] implemented this method for large datasets, and was extended to CLARANS[58] for spatial data mining.

Later on the more robust trimmed k-means method has been proposed[59], inspired by the trimming ideas in the MCD and the LTS. It searches for the h-subset (with h as in the definition of MCD) such that the sum of the squared distances of the observations to the mean of their group is minimized. Consequently, not all observations need to be classified, as $n-h$ cases can be left unassigned. To perform the trimmed k-means clustering an iterative algorithm[60] has been developed, using C-steps like those in the FAST-MCD algorithm. For non-spherical clusters, constrained maximum likelihood approaches[61,62] were developed.

Functional data

In functional data analysis, the cases are not data points but functions. A functional data set typically consists of n curves observed on a set of gridpoints $t_1, \ldots, t_T$. These curves can have smoothness properties, numerical derivatives and so on. Standard references on functional data are the books[63,64]. A functional data set can be analyzed by principal components, for which robust methods are available[65]. To classify functional data, a recent approach is presented in[66].

The literature on outlier detection in functional data is rather young, and several graphical tools have been developed[67–69], mainly for univariate functions. In[70] also multivariate

functions are discussed and a taxonomy of functional outliers is set up, with on the one hand functions that are outlying on most of their domain, such as shift and magnitude outliers as well as shape outliers, and on the other hand isolated outliers which are only outlying on a small part of their domain. The proposed heatmap and functional outlier map are tools to flag outliers and detect their type. This work is expanded in[20] to functional data with bivariate domains, such as images and video.

Other applications

Robust statistics has many other uses apart from outlier detection. For instance, robust estimation can be used in automated settings such as computer vision[71,72]. Another aspect is statistical inference, such as the construction of robust hypothesis tests, p-values, confidence intervals, and model selection (e.g. variable selection in regression). This aspect is studied in[3,7] and in other works they reference.

DETECTING OUTLYING CELLS

Until recently people have always considered outliers to be cases (data points), i.e. rows of the $n \times d$ data matrix $\boldsymbol{X}$. But recently the realization has grown that this paradigm is no longer sufficient for the high-dimensional data sets we are often faced with nowadays. Typically most data cells (entries) in a row are regular and a few cells are anomalous. The first paper to formulate the cellwise paradigm was[73], which showed how such outlying cells propagate in computations. In more than a few dimensions, even a small percentage of outlying cells can spoil a large percentage of rows. This is fatal for rowwise robust methods, which require at least 50% of the rows to be clean.

Detecting cellwise outliers is a hard problem, since the outlyingness of a cell depends on the relation of its column to the other columns of the data, and on the values of the other cells in its row (some of which may be outlying themselves). The *DetectDeviatingCells*[74] algorithm addresses these issues, and apart from flagging cells it also provides a graphical output called a cellmap.

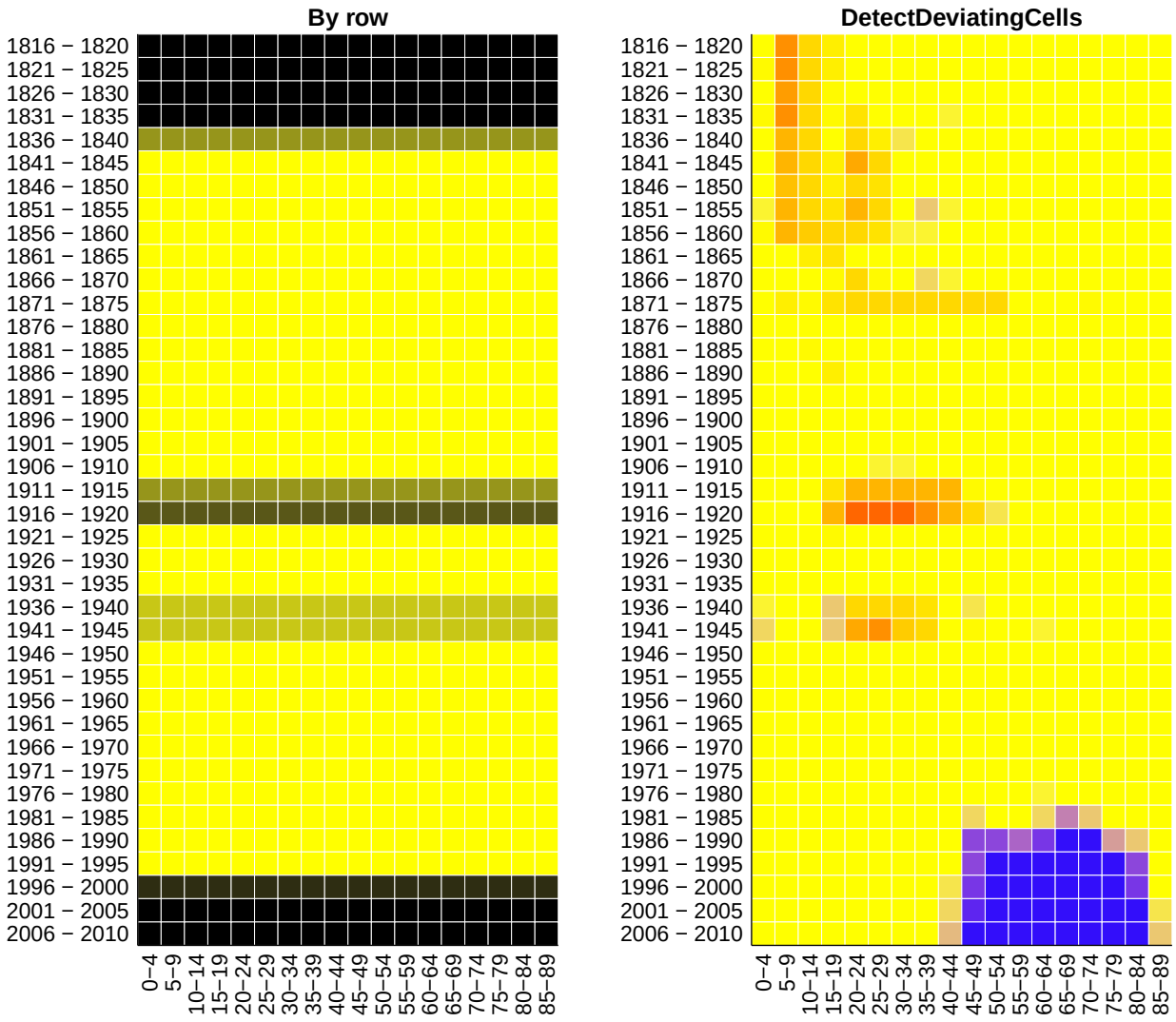

Figure 8: Male mortality in France in 1816–2010: (left) detecting outlying rows by a robust PCA method; (right) detecting outlying cells by *DetectDeviatingCells*. After the analysis, the cells were grouped in blocks of 5×5 for visibility.

As an example we consider the mortality by age for males in France from 1816 to 2010, obtained from http://www.mortality.org . Each row corresponds to the mortalities in a given calendar year. The left panel in Figure 8 shows the result of the ROBPCA method described in the section on principal components. Outlying rows are shown in black and regular rows in yellow. The analysis was carried out on the data set with the individual years and the individual ages, but as this resolution would be too high to fit on the page we

have combined the cells into 5×5 blocks afterward. The combination of some black rows with some yellow ones has led to gray blocks. We can see that there were outlying rows in the early years, the most recent years, and during two periods in between. Note that a black row doesn't provide information about its cells.

By contrast, the result of *DetectDeviatingCells* in the right panel in Figure 8 identifies a lot more information. Cells with higher values than predicted are shown in red, and those with lower values in blue, after which the colors were averaged in the 5×5 blocks. The outlying early years saw a high infant mortality. During the Prussian war and both world wars there was a higher mortality among young adult men. And in recent years mortality among middle-aged and older men has decreased substantially, perhaps due to medical advances.

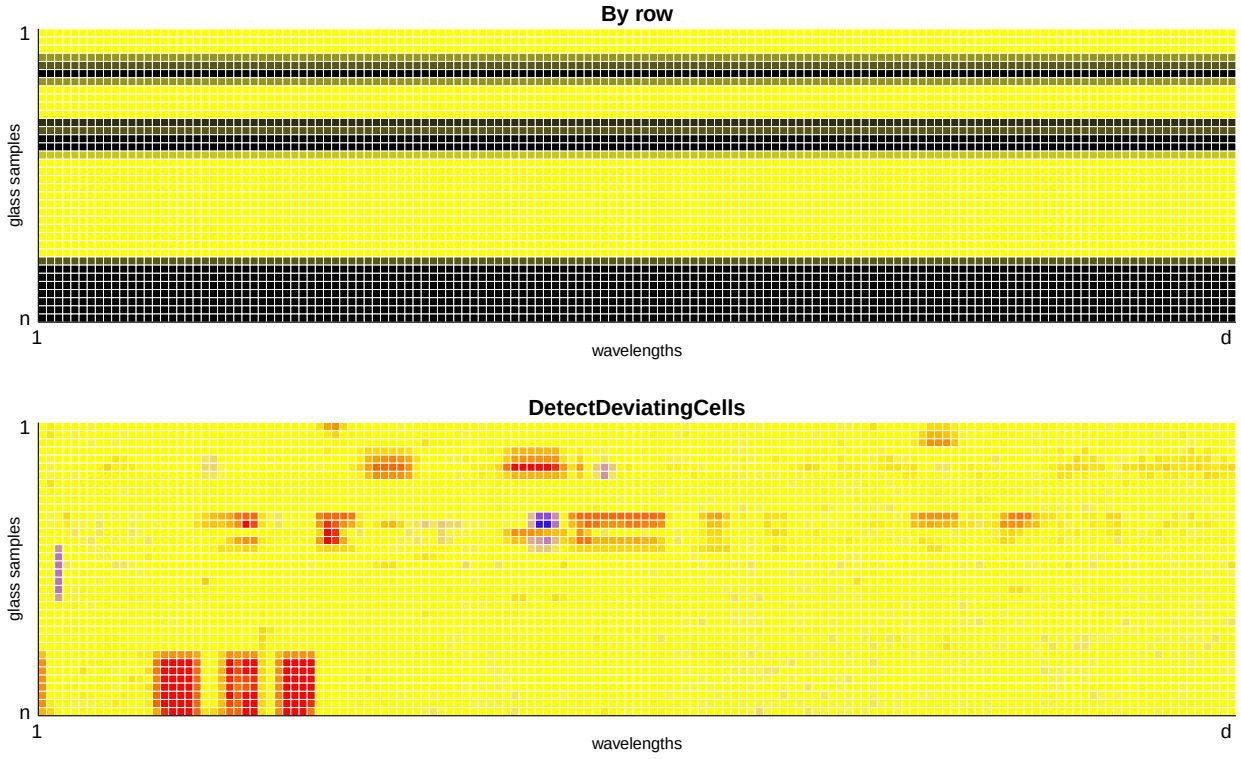

Figure 9: Cell map of the glass data. The positions of the deviating cells reveal the chemical contaminants.

We also return to the glass data from the section on PCA. The top panel in Figure 9 shows the rows detected by the ROBPCA method. The lower panel is the cell map obtained

by *DetectDeviatingCells* on this 180×750 dataset. After the analysis, the cells were again grouped in 5×5 blocks. We now see clearly which parts of each spectrum are higher/lower than predicted. The wavelengths of these deviating cells reveal the chemical elements responsible.

Ideally, after running*DetectDeviatingCells* the user can look at the deviating cells and whether their values are higher or lower than predicted, and make sense of what is going on. This may lead to a better understanding of the data pattern, to changes in the way the data are collected/measured, to dropping certain rows or columns, to transforming variables, to changing the model, and so on. If the data set is too large for visual inspection of the results, or the analysis is automated, the deviating cells can be set to missing after which the data set is treated by a method appropriate for incomplete data. A good rowwise robust method of this type is[75].

SOFTWARE AVAILABILITY

All the examples in this paper were produced with the free software R[76]. The publicly available CRAN package `robustbase` contains Qn, covMcd, ltsReg, and lmrob, whereas `rrcov` has many robust covariance estimators, robust principal components, and robust LDA and QDA classification. ROBPCA and its extensions are available in `rospca` and robust SIMCA in `rrcovHD`. Robust clustering can be performed with the `cluster` and `tclust` packages. The package `mrfDepth` contains tools for functional data and `cellWise` provides cellwise outlier detection and cellmaps.

Matlab functions for many of these methods are available in the LIBRA toolbox[77,78], which can be downloaded from `http://wis.kuleuven.be/stat/robust` .

The MCD and LTS methods are also built into S-PLUS as well as SAS (version 11 or higher) and SAS/IML (version 7 or higher).

CONCLUSIONS

We have surveyed the utility of robust statistical methods and their algorithms for detecting anomalous data. These methods were illustrated on real data, in frameworks ranging from covariance matrices, the linear regression model and principal component analysis, with references to methods for many other tasks such as supervised and unsupervised classification as well as the analysis of functional data. For high-dimensional data, sparse and regularized robust methods were developed recently.

We have described methods to detect anomalous cases (rowwise outliers) but also newer work on the detection of anomalous data cells (cellwise outliers). An important topic for future research is to further improve the efficiency of the robust methodologies, in terms of both predictive accuracy and computational cost.

References

1. M. Hubert, P.J. Rousseeuw, and S. Van Aelst. High breakdown robust multivariate methods. *Statistical Science*, 23:92–119, 2008.

2. P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. Wiley-Interscience, New York, 1987.

3. R.A. Maronna, D.R. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods*. Wiley, New York, 2006.

4. F.R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42:1887–1896, 1971.

5. D.L. Donoho and P.J. Huber. The notion of breakdown point. In P. Bickel, K. Doksum, and J.L. Hodges, editors, *A Festschrift for Erich Lehmann*, pages 157–184, Belmont, 1983. Wadsworth.

6. M. Hubert and M. Debruyne. Breakdown value. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1:296–302, 2009.

7. F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, 1986.

8. P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283, 1993.

9. P.J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

10. M. Hubert and E. Vandervieren. An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, 52:5186–5201, 2008.

11. G. Brys, M. Hubert and A. Struyf. A robust measure of skewness. *Journal of Computational and Graphical Statistics*, 13:996–1017, 2004.

12. P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

13. P.J. Rousseeuw. Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, editors, *Mathematical Statistics and Applications, Vol. B*, pages 283–297, Dordrecht, 1985. Reidel Publishing Company.

14. M. Hubert and M. Debruyne. Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:36–43, 2010.

15. P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics*, 41:212–223, 1999.

16. M. Hubert, P.J. Rousseeuw, and T. Verdonck. A deterministic algorithm for robust location and scatter. *Journal of Computational and Graphical Statistics*, 21:618–637, 2012.

17. K. Boudt, P.J. Rousseeuw, S. Vanduffel, and T. Verdonck. The Minimum Regularized Covariance Determinant estimator. Technical Report, *arXiv:1701.07086*, 2017.

18. W.A. Stahel. *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. PhD thesis, ETH Zürich, 1981.

19. D.L. Donoho. Breakdown properties of multivariate location estimators. Ph.D. Qualifying paper, Dept. Statistics, Harvard University, Boston, 1982.

20. P.J. Rousseeuw, J. Raymaekers, and M. Hubert. A measure of directional outlyingness with applications to image data and video. *Journal of Computational and Graphical Statistics*, to appear. arXiv:1608.05012, 2016.

21. R.A. Maronna and V.J. Yohai. The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90:330–341, 1995.

22. R.A. Maronna. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics*, 4:51–67, 1976.

23. L. Davies. Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15:1269–1292, 1987.

24. K.S. Tatsuoka and D.E. Tyler. On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *The Annals of Statistics*, 28:1219–1243, 2000.

25. R.A. Maronna and R.H. Zamar. Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics*, 44:307–317, 2002.

26. P.J. Rousseeuw and K. Van Driessen. Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12:29–45, 2006.

27. G. Pison, S. Van Aelst and G. Willems. Small sample corrections for LTS and MCD. *Metrika*, 55:111–123, 2002.

28. G. Willems and S. Van Aelst. Fast and robust bootstrap for LTS. *Computational Statistics and Data Analysis*, 48:703–715, 2005.

29. M. Salibian-Barrera, S. Van Aelst and G. Willems. Fast and robust bootstrap. *Statistical Methods and Applications*, 17:41–47, 2008.

30. P.J. Rousseeuw and B.C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–651, 1990.

31. K.A. Brownlee. *Statistical Theory and Methodology in Science and Engineering*. Wiley, New York, 1965.

32. J. Khan, S. Van Aelst and R. Zamar. Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics and Data Analysis*, 52:239–248, 2007.

33. J. Khan, S. Van Aelst and R. Zamar. Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102:1289–1299, 2007.

34. A. Alfons, C. Croux and S. Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7:226–248, 2013.

35. P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

36. J. Jurecková. Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, 42:1328–1338, 1971.

37. R. Koenker and S. Portnoy. L-estimation for linear models. *Journal of the American Statistical Association*, 82:851–857, 1987.

38. P.J. Rousseeuw and V.J. Yohai. Robust regression by means of S-estimators. In J. Franke, W. Härdle, and R.D. Martin, editors, *Robust and Nonlinear Time Series Analysis*, pages 256–272, New York, 1984. Lecture Notes in Statistics No. 26, Springer-Verlag.

39. V.J. Yohai. High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15:642–656, 1987.

40. C. Croux and G. Haesbroeck. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87:603–618, 2000.

41. M. Salibian-Barrera, S. Van Aelst, and G. Willems. PCA based on multivariate MM-estimators with fast and robust bootstrap. *Journal of the American Statistical Association*, 101:1198–1211, 2006.

42. M. Hubert, P.J. Rousseeuw, and S. Verboven. A fast robust method for principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60:101–111, 2002.

43. C. Croux, P. Filzmoser, and M.R. Oliveira. Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87:218–225, 2007.

44. M. Hubert, P.J. Rousseeuw, and K. Vanden Branden. ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47:64–79, 2005.

45. P. Lemberge, I. De Raedt, K.H. Janssens, F. Wei, and P.J. Van Espen. Quantitative Z-analysis of 16th-17th century archaeological glass vessels using PLS regression of EPXMA and μ-XRF data. *Journal of Chemometrics*, 14:751–763, 2000.

46. N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, and K.L. Cohen. Robust principal component analysis for functional data. *Test*, 8:1–73, 1999.

47. M. Hubert, T. Reynkens, E. Schmitt, and T. Verdonck. Sparse PCA for high-dimensional data with outliers. *Technometrics*, 58:424–434, 2016.

48. I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

49. M. Debruyne and T. Verdonck. Robust kernel principal component analysis and classification. *Advances in Data Analysis and Classification*, 4:151–167, 2010.

50. D.M. Hawkins and G.J. McLachlan. High-breakdown linear discriminant analysis. *Journal of the American Statistical Association*, 92:136–143, 1997.

51. X. He and W.K. Fung. High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72:151–162, 2000.

52. C. Croux and C. Dehon. Robust linear discriminant analysis using S-estimators. *The Canadian Journal of Statistics*, 29:473–492, 2001.

53. M. Hubert and K. Van Driessen. Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, 45:301–320, 2004.

54. S. Wold. Pattern recognition by means of disjoint principal components models. *Pattern Recognition*, 8:127–139, 1976.

55. K. Vanden Branden and M. Hubert. Robust classification in high dimensions based on the SIMCA method. *Chemometrics and Intelligent Laboratory Systems*, 79:10–21, 2005.

56. J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

57. L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

58. R.T. Ng and J. Han. CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14:1003–1016, 2002.

59. J.A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25:553–576, 1997.

60. L.A. García-Escudero, A. Gordaliza, and C. Matrán. Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics*, 12:434–449, 2003.

61. M.T. Gallegos and G. Ritter. A robust method for cluster analysis. *The Annals of Statistics*, 33:347–380, 2005.

62. L.A. García-Escudero, A. Gordaliza, C. Matrán, and A. Mayo-Iscar. A general trimming approach to robust cluster analysis. *The Annals of Statistics*, 36:1324–1345, 2008.

63. J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

64. F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York, 2006.

65. G. Boente and M. Salibian-Barrera. S-estimators for functional principal component analysis. *Journal of the American Statistical Association*, 110:1100-1111, 2015.

66. M. Hubert, P. Rousseeuw, and P. Segaert. Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, to appear. doi: 10.1007/s11634-016-0269-3.

67. R. Hyndman and H. Shang. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19:29–45, 2010.

68. Y. Sun and M. Genton. Functional boxplots. *Journal of Computational and Graphical Statistics*, 20:316–334, 2011.

69. A. Arribas-Gil and J. Romo. Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15:603–619, 2014.

70. M. Hubert, P.J. Rousseeuw and P. Segaert. Multivariate functional outlier detection (with discussion). *Statistical Methods & Applications*, 24:177–246, 2015.

71. P. Meer, D. Mintz, A. Rosenfeld, and D.Y. Kim. Robust regression methods in computer vision: a review. *International Journal of Computer Vision*, 6:59–70, 1991.

72. C.V. Stewart. MINPRAN: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:925–938, 1995.

73. F. Alqallaf, S. Van Aelst, S., V.J. Yohai, and R.H. Zamar. Propagation of outliers in multivariate data. *The Annals of Statistics*, 37:311–331, 2009.

74. P.J. Rousseeuw and W. Van den Bossche. Detecting deviating data cells. *Technometrics*, to appear. arXiv:1601.07251, 2016.

75. C. Agostinelli, A. Leung, V.J. Yohai, and R.H. Zamar. Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination. *Test*, 24:441–461, 2015.

76. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria, https://www.R-project.org/, 2016.

77. S. Verboven and M. Hubert. LIBRA: a Matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75:127–136, 2005.

78. S. Verboven and M. Hubert. MATLAB library LIBRA. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2: 509–515, 2010.